DOCUMENT RESUME

ED 206 690 TM 810 589

AUTHOR Fremer, John
TITLE Impact of Test Disclosure Legislation on Test Development.
PUB DATE Apr 81
NOTE 17p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Equated Scores; *State Legislation; *Test Construction; *Testing Programs
IDENTIFIERS Educational Testing Service; Graduate Management Admission Test; Graduate Record Examinations; Law School Admission Test; New York; Preliminary Scholastic Aptitude Test; Scholastic Aptitude Test; Test Disclosure; *Testing Industry; Test of English as a Foreign Language; *Truth in Testing Legislation

ABSTRACT

Test disclosure legislation in New York State (LaValle Act) has had a major impact on the national testing programs administered by Educational Testing Services (ETS) for various sponsoring organizations. The paper reviews the immediate operational effects of test disclosure in the following areas: (1) increase in number of test forms developed; (2) acceleration of development of new equating methods; and (3) filing requirements and interpretive materials. Possible future changes in national testing programs also discussed are: impact of altering one of the basic conditions of testing; broadening input to the testing process; and changes in the nature of tests. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206690

Submitted to Educational Researcher - June 1981 (slightly revised version - June 26, 1981)

Impact of Test Disclosure Legislation on Test Development[1,2]

John Fremer

Educational Testing Service

Princeton, New Jersey

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✕ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Fremer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

OUTLINE OF PAPER

A. Introduction

B. Immediate Operational Effects on Test Programs

- Increase in number of test forms developed
- Acceleration of development of new equating methods
- Filing requirements and interpretive materials

C. Groundwork for Future Changes

- Altering one of the basic conditions of testing programs
- Broadening input to the test development process
- Possible changes in the nature of tests

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981. Part of Symposium 30.03 -- "Test Disclosure Legislation -- Issues and a First Look at the Evidence."

[2]The perspective represented in this paper is that of the author. The paper does not represent a policy statement either for Educational Testing Service or for any of the organizations which sponsor the testing programs that are discussed.

TM 810 589

INTRODUCTION

This paper reviews the initial impact of the LaValle Test Disclosure Act on six testing programs administered by Educational Testing Service for various sponsoring organizations. Table 1 lists the testing programs, sponsors, and program purposes.

The tests that are discussed include two that are taken by high school students:

Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT)

Scholastic Aptitude Test (SAT)

There is one test that may be taken either by high school or college level students applying for admission to post-secondary education in the U.S.

Test of English as a Foreign Language (TOEFL)

There are three tests for students in college or college graduates who are planning to go on for further academic and professional education:

Graduate Management Admission Test (GMAT)

Graduate Record Examinations Aptitude Test (GRE)

Law School Admission Test (LSAT)

Testing Programs Administered by Educational Testing Service

Affected by New York State Test Disclosure Law

(LaValle Act)

Examinations	Program Sponsors	Purposes of Examinations
1. Graduate Management Admission Test (GMAT)	Graduate Management Admission Council	Serves role in admission to graduate study in management
2. Graduate Record Examination Aptitude Test (GRE)	Graduate Record Examinations Board	Serves role in admission to graduate study in arts and sciences
3. Law School Admission Test (LSAT)	Law School Admission Council	Serves role in admission to study of law
4. Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT)	College Board, National Merit Scholarship Corporation	Serves role in guidance and as initial screen for National Merit Scholarship Program
5. Scholastic Aptitude Test (SAT)	College Board	Serves role in admission to undergraduate study
6. Test of English as a Foreign Language (TOEFL)	College Board, Graduate Record Examinations Board, and Educational Testing Service	Serves role in both undergraduate and graduate admissions

IMMEDIATE OPERATIONAL EFFECTS ON TESTING PROGRAMS

Increase in Number of Test Forms Developed

Prior to the passage of the LaValle Act in New York State it was possible to offer tests such as the Scholastic Aptitude Test (SAT) and the Graduate Record Examinations Aptitude Test (GRE) on many different occasions without building a new test form for each occasion. The individual test forms remained "secure". No student received copies of the test questions after the administration. Sample tests were available to students, however, along with explanations of the test content, test development process, and purposes for the tests.

Secure testing programs could maintain a small inventory of test forms, adding a small number of new forms each year, and retiring from use the oldest versions. The amount of money devoted to operational test development was a small part of the program budget. Additional funds were used to support research on the design, development, and validation of new types of questions.

Once the LaValle Act took effect in New York State on January 1, 1980 any test covered by the legislation had to become public thirty days after scores were reported. In order to preserve the small available inventory of test forms, test program sponsors took such actions as reducing the number of test administrations in New York State. To maintain the testing programs over time, however, increases in test development had to be initiated.

The two programs requiring the smallest increase in development were the PSAT/NMSQT and TOEFL. Paradoxically, these were the programs with the smallest and largest volumes of test development prior to disclosure. The

PSAT/NMSQT, which is administered on only two dates in a single week in October, continued to need only two new test forms annually. TOEFL, which is administered monthly in states other than New York State, was developing many test forms per year even before passage of the LaValle Act. This high rate of test development for TOEFL was needed to maintain an international testing program under secure conditions. Both the PSAT/NMSQT and TOEFL, therefore, needed only small increases in test development work. Most of this additional work was related to the preparation of interpretive materials to comply with LaValle. Information already contained in student bulletins and test manuals had to be organized into appropriate supplementary documents.

For the SAT, GMAT, GRE, and LSAT the increases in workload were quite significant: There was about a 45% increase for the SAT as 10 new forms a year were necessary instead of the seven that had been developed previously. The increase for the GMAT, GRE, and LSAT was 100 to 150%, as more than twice as many new examinations as were needed in the past had to be developed each year.

This increase in admissions-related test development, moreover, occurred at the same time that other changes in the design of some of the tests were being required because of the impact of disclosure on test equating. It was necessary to add new test development staff and to increase arrangements for outside help for test question writing and review. Major burdens were placed on experienced staff in order to meet the increase in new forms needed. Inevitably, the programs affected by disclosure drew large amounts of staff time away from other ongoing development activities. It would be difficult to overstate the amount of disruption that resulted.

Acceleration of Development of New Equating Methods

The LaValle Act required disclosure for all items contributing to a student's test score but excluded items included in a test for reasons of pretesting and equating. The impact of the legislation varied from program to program depending on the method of equating used previously. Table 2 summarizes the equating procedures employed by the six programs prior to disclosure and indicates the effect of the new legislative requirement.

The SAT used separate anchor test equating. Since anchor tests do not contribute to the scores of students, these equating subtests tests can remain secure and the method can still be used.

TOEFL used Item Response Theory Equating. Items can be precalibrated before they contribute to students' scores. The precalibrated items can be used only once for equating purposes since they contribute, at that point, to the scores of students. The method can continue to be used, however, under test disclosure conditions.

Table 2

Effect of LaValle Act on Equating Method Used

Prior to Disclosure for Six Testing Programs Administered by ETS

Examination	Equating Method Used	Possible to Continue Method?
SAT	Separate Anchor Test	Yes -- No problem
TOEFL	Item Response Theory	Yes -- No problem
PSAT/NMSQT	Embedded Common Items	For a while -- as long as undisclosed old SAT forms are available
GMAT, GRE Aptitude, and LSAT	Spiralling -- New form given with one or more old forms	No -- Old and new forms will be disclosed

The PSAT/NMSQT was equated through embedded common items from retired forms of the SAT. Since these items do contribute to the scores of students, the method can be used only as long as undisclosed forms of the SAT are available.

For the GRE Aptitude Test, GMAT, and LSAT the method of equating that had been employed is called "spiralling". Spiralling requires the administration along with the new test, of one or more old test forms that are already on the test score scale. At a GRE Aptitude administration where a new test form was being introduced, one-half the students might take this new test and one-half an old test. Since tests were assigned randomly to students, differences between the raw scores on the two tests could be attributed to the differences in the relative difficulty of the tests and appropriate equating adjustments made. Under this model, though, the entire old test must remain secure. With disclosure one can use a test only once. Then it must be disclosed and thus rendered unusable for future equating. For the GRE, GMAT, and LSAT, therefore, the LaValle Act struck down the existing method of equating.

Clearly alternative equating models had to be employed. In the case of the GRE the program continued to employ the spiralling method at nondisclosed administrations in states other than New York, but also moved to develop sets of separate anchor tests for each of its three scores -- analytical, mathematical, and verbal. (These tests represented additional test development work and expense in a program where the staff involved were also faced with the additional work required to increase total test form production.) The format of the test was changed to equalize section

timing and the total test time was increased so that equating sections could be administered routinely as part of the program.

Reducing the number of administration dates in New York State and creating the potential for anchor test equating provided the GRE program with the time needed to plan and carry out additional experimentation. Nondisclosed administrations can be used to compare the results obtained by spiralling, anchor test equating, and the use of item response theory equating.

The choice of an alternative equating method for the GMAT and LSAT is a more difficult problem. The heterogeneous nature of these examinations argues against the use of either anchor test equating or item response theory equating until additional experimental work is carried out. Research is needed extending across several administrations and candidate samples. Although research or IRT equating for these tests has been proposed, it could not be carried out within the developmental schedule needed to insure compliance with the LaValle legislation.

For the GMAT and LSAT, as well as other testing programs, the program sponsors and ETS have the responsibility either to use psychometric methods that assure that scores from different test forms have the same meaning or, if there must be a break in the continuity of the score scale, to alert score users about the change in the meaning of scores. In either case steps must be taken to avoid misinterpretation of test scores. Breaking the score scale for a program may be necessary to permit the introduction of a new equating method, but that method or a comparable one must then be used to place future test forms on the new score scale.

For both the GMAT and LSAT we are currently experimenting with a newly developed procedure known as section preequating. This method requires the administration of a final test experimentally in a series of pairs of sections prior to the operational use of the test. Since section preequating experimentation has had to take place at the same time as a major pretesting effort for the large amount of newly written test questions, as many as 40 different versions of test sections that do not count toward the reported score have had to be included with the operational test used at a particular administration. The logistics of developing such a complex system and seeing it through production, administration, and analysis have been formidable.

As part of the adaptation of the GMAT to the section pre-equating method, the test has been reformatted so that each section has the same timing. The total time for the GMAT has also been increased, initially by one-half hour, but with a future time increase also to be required. The total increase in time will be 55 minutes.

For the LSAT more significant changes in test content are being undertaken and the program sponsor, the Law School Admission Council, has decided to introduce a new score scale. The use of a new score scale will make it clear to score users that the nature of the test has undergone a major change. Scores on the new score scale will not be linked to scores on the old scale.

Although the development of a new equating method ought to have years of research to confirm its effectiveness and explore its implications, the LaValle Act has greatly reduced the lead time available for such work. While the experimentation underway will almost certainly hasten the use of

new equating tools, it is hard to ignore the potential dangers of having to follow a developmental timetable that is based on a legislative mandate. A mandate, moreover, that makes little provision for the research and analysis that is consistent with high measurement standards.

Filing Requirements and Interpretive Materials

The LaValle legislation required the filing of background reports and statistical data regarding the affected tests with the Commissioner of the State of New York. In the debate over the need for the LaValle legislation, ETS and other test publishers argued that high quality technical and interpretive material was readily available, while critics of testing contested this claim. (See Strenio, 1979, ACT, 1980, and Brown, 1980.)

The 1979 LaValle Act was ammended in 1980 in a manner that protected institutional privacy and which provided some flexibility as to the point in the testing process at which information could be provided to candidates. It proved possible for ETS to meet the filing requirements of LaValle through the submission of materials already available to interested parties prior to the legislation. Reports from New York State indicate that the large amount of materials that is now on file with the Department of Education has received little use. For example, there appear to have been no inquiries to Educational Testing Service during the first 18 months of the LaValle Act (January 1, 1980 to June 30, 1981) that are related to the technical materials filed with the Commissioner.

In addition to filing materials with the Commissioner of Education in Albany, New York, the test program sponsors provided additional material

to test takers who requested copies of their test booklets. In most instances the material accompanying the booklet was relatively brief, providing, for example, an explanation of the procedure used to derive a scaled score.

In the TOEFL program an additional package of materials "Understanding TOEFL: Test Kit 1," going beyond the requirements of the legislation, was also prepared. The Test Kit contains a complete TOEFL test along with an explanation of each question in the test and of the four answer choices for each question. The very large response to the Test Kit makes clear that this publication is filling an important information need.

GROUNDWORK FOR FUTURE CHANGES

Altering One of the Basic Conditions of Testing Programs

The greatest impact of test disclosure at ETS may stem from the change in attitude it is bringing about in the contributors to test development for national admissions testing programs. The concept of test security has been central to many aspects of the development of major testing programs that are providing information for individuals and institutions. Now security of test questions can be preserved only up to the time of the first operational use of a test.

The initial accommodations to disclosure that are described earlier in this paper permit testing to continue while more attention is given to additional alternatives. Once test development and equating reach a stable state, it is likely that the entire test development and analysis process at ETS will be analyzed and redesigned. Major change is predictable because test disclosure, although a major force, is not the only pressure

for innovation on the test development process. Advances in technology and in measurement theory also push on our current methods. As a result, at Educational Testing Service, groups of staff representing different areas of specialization and interests are collaborating on reviews of our most tradition-bound assumptions. Fundamental changes in measurement practice appear to be both possible and desirable.

Broadening Input to the Test Development Process

One theme that cuts across many of the emerging trends in test development at ETS is that of additional external involvement. The participation of educators from schools and colleges in the development of examinations has long been a standard feature of our work. There is increasing evidence, though of ETS and program sponsor interest in broadening still further the role played by such external contributors. Such actions are viewed as ways of demonstrating our commitment to public accountability and our openness to the ideas of those impacted by testing.

Among the kinds of activities that support this generalization are the following:

- An increase in the number of opportunities for students to sit with subject-matter committees to discuss the experience of taking a test that had been developed with the help of the committee.
- The establishment by test program sponsors of Test Question Review Committees with responsibility for reviewing each question before it appears on a final test. (A practice already in place for many major tests prior to test disclosure legislation.)
- The development of a specifications survey model to provide additional input into the determination of test content specifications, and

to evaluate the match between test questions and the associated specifications

- The formation, by the ETS Senior Vice President for Testing Programs, of an external advisory group. This group includes among its members representatives of student associations, representatives of other educational groups, and individuals prominent in educational research and measurement.

External input to ETS test development includes letters from test takers and the media coverage given to our work. In some instances our activities have even moved from no news coverage at all to the education pages of newspapers to page one. Such attention does provide an opportunity to explain what we believe to be the strengths and weaknesses of our tests to quite a wide audience. If the result is that readers and listeners end up with a balanced picture, it may be all to the good. Former ETS President William Turnbull noted that educational testing is in greater danger from the zealots who perpetuate fallacies, such as the notion that tests measure with infallible precision, than from our critics (Turnbull, 1978). Anyone who is not willing to take strong criticism has no business working in a field where improvement through criticism is so fundamental.

Possible Changes in the Nature of Tests

It is my hope that the thorough shaking which test disclosure has given to testing and the continuation of the trend toward greater external involvement will result in a number of desirable modifications to testing. My reading of the kinds of input we are getting suggests that we will be facing increasingly greater pressure not only to pr evidence of validity for our tests for their intended purposes but also to justify

each question that we include on a test. It will not be enough to explain the steps leading to specifications calling for a question of a particular type. We will be called upon to provide evidence that the test questions measure the construct that the test is designed to measure.

Not only do I expect attention to extend to individual test questions but also to the answer choices for these questions. What is the basis for giving credit to a particular answer in scoring? How do the test directions, the descriptive material, and the question itself establish the context for identifying that response as the key? In the past, in highly structured areas such as mathematics we have argued the self evident "correctness" of the keyed answer to those with the appropriate mathematical skill. Experiences with question challenges, though, make it clear that variations in the interpretation of language can exist in any discipline. It is the test development process itself that must serve as the basis for accepting as appropriate a test with its associated key. To reach that end, the process must permit the test taker and other interested people a chance to "challenge the experts" and be heard.

Despite my interest, and that of my ETS colleagues, in attempting to extract some benefit in the future for education and testing from disclosure, I see some further negative outcomes looming ahead. I believe, for example, that the 1980 extension of the LaValle Act to cover achievement examinations is likely to result in the complete elimination, in all states, of certain of these examinations. As of April 1981, only six of the twenty GRE Advanced tests are offered in New York State. The loss of New York testing volume for some of the remaining tests that already have small volumes further reduces their economic viability. Since small volume

tests also pose problems in obtaining equating samples, at a minimum I expect to see a reduced number of offerings nationally.

Another possible negative consequence of disclosure is a shift in test question content away from more imaginative and insightful types and towards conventional forms. This type of shift could occur in an effort to avoid questions that are susceptible to criticism. At worst we could end up with tests of mathematical skills that emphasized what one of my ETS colleagues calls rat-tat-tat mathematics. The test development staff at ETS are working with external contributors in an attempt to develop procedures that will permit the test development enterprise to continue to evolve in positive directions without being trapped into the kind of content limitations that are a potential danger.

Closing Comment

This paper calls attention to the major impacts on test development volumes and equating procedures at ETS as a result of test disclosure. It is possible that the public attention that testing is receiving and the increase in external assistance in test planning and development will lead to more positive future effects. As an ETS test developer I believe that I need to listen carefully to the many kinds of suggestions that are being made and to be willing to challenge past assumptions. As I see it neither the critics nor the proponents of testing are urging abandonment of high measurement standards or of commitment to fairness in testing and test use. What is being asked for, though, is a willingness to explore alternative routes to these same ends.

REFERENCES

American College Testing Program. Standardized Testing Legislation. American College Testing Program, 1980.

Brown, Rexford. Searching for the Truth about "Truth in Testing" Legislation. Denver, Colorado: Education Commission of the States, January 1980.

Strenio, Andrew Jr. The Debate Over Open Versus Secure Testing: A Critical Review. National Consortium on Testing, Staff Circular No. 6, Cambridge, Massachusetts, Huron Institute, October 1979.

Turnbull, William. Test scores in perspective. Paper presented at the National Conference on Achievement Testing in the Basic Skills, Washington, D.C., March 1978.